John O'Rourke
Senior Research Proposal Draft 3

Automating a User Defined Categorization of the Web

Abstract:

Categorizing the contents of the web is presently done manually by human "librarians" who visit a website and then define under which topic headings the web page or site belongs. This is a tedious and expensive process done by many portal websites. Research has been and is being done to try to develop a system that automates this process of Web Categorization. Present research techniques include study of human retrieval behavior, developing a system that limits/automates the interaction between the client the "librarian", the clustering similar websites and using metadata tags to describe keywords, and other information placed inside the html code itself. These processes are insufficient because they lack knowledge of what features are included in particular categories and how the user may choose to search for a specific site on the web. A present solution to this problem is to allow the user of the categorization tool to define their own specific categories and information about those specific categories on the web. This would allow a more clearly defined way to automate the categorization of websites/pages.

Introduction:

There is a need to automatically categorize the web. Today, the present search engine sites such as Yahoo categorize websites by having a human visit the site and decide which categories are appropriate to be declared as categories linking to the website [1]. This process is both a costly and time consuming. The labor costs are tremendous and the size of the web is continuing to grow at a rapid pace leaving more and more sites to be categorized by hand each day. Without an automatic categorization system the result of a website changing its content completely will result in an inaccurate categorization. Once the site's content has been recognized by a human to have been changed the "librarian" must reexamine the website and categorize it based upon the new evaluation. Thus, a need for automatic categorization is a necessity to assist the average

user to find the website that he/she is looking for and to reduce time consuming labor costs that will never cease to exist for search engine such a Yahoo [1].

Background:

In order to understand how to group web pages one must first observe how a client interacts with a search engine. Most average clients search using previous experiences in searching. If the Internet is a library for example, the first task a user wishes to perform is to view a catalogue of what is inside the library. In order to create this catalogue an association is required between certain books in the library. Once the user has selected some books, other similar books may be searched for based upon the users‘ responses to the books returned. This would basically be the observation and similarity of books according the users personal preferences in what they wish to find [7].

An application must act as both a sorting mechanism for the Web and a librarian for the user. In order to evaluate the effectiveness sorting and retrieving is the percentage of relevant pages that are returned based upon each query. This is a weighted ranking of each site according to the search specified by the user. One way to assure of this is have a pop up list of terms for the user to select when he/she has entered a query, this allows the user to modify his/her query to in order to conform to information that is relevantly recognized by the database [8].

The concept of clustering similar websites together by evaluating the html code itself is a possible way to “card catalog” the web. In this scenario no prior knowledge is needed in order to group similar websites together. By evaluating relevant data in the html document a general idea about the document may be retrieved. The first step in this process is to extract extraneous data from the html site, leaving only significant information to be evaluated on a particular website. Once this is done a word evaluation occurs in which the frequency of the words is counted. At this point link information

may also be retrieved and stored as well. This information may be used to find other sites that are not already listed in the database. For example if the database contains two web URLs whose pages have links to a third site, most likely the third site most likely will contain the same subject as the previous two. Once this information has been gathered similar websites may be associated with it according to the criteria stated above. In theory if a query is entered searching for a particular type of website all of the websites that have similarities according to words and linking URLs are returned as hits for that particular search [2][3].

Another attempt to "card catalog" the web is to place metadata tags describing the contents of particular websites in the html documents themselves. A meta tag is an html tag that is not used to generate an html page but is contained in the html document and used to describe a website. Presently authors of a website must insert possible categories and keywords that they wish to link to their site. A categorization then takes place when a site such as Yahoo evaluates the website. When a web site which contains these meta tags is evaluated it will be placed in the categories and keywords defined by the html author. This puts more burden on the creator of the website and less labor intensive classifying on the behalf of web portals [4].

These methods attempt to create a computerized "library catalog" and "librarian" but they actually create a group of "librarians", "binary catalog" and different "catalog" for every library. The concept of creating pop-up options for the user is lengthy and inconvenient. Creating a forced pop up list is the same as going to the library and having five different librarians all describing different ideas at the same time. The list is unclear

and a hassle for the user. What would be more beneficial is an option to be clicked that would suggest alternate queries that the user may type in.

A “binary catalog” is created when the clustering concept is implemented. This means that the only thing that can understand the clustering is a computer. In other words a user cannot just browse through an English(as compared to Binary) version of categories in order to find what one is looking for. The user must instead always enter a query in order to search the web. Also the clustering idea generally can group similar websites but has in the past had limited word storage capabilities. Banners also get in the way of link evaluation accuracy. Because there is no prior knowledge given about the information stored in the database the clusters could be anything. There is not a common human phrase to describe each cluster, sure there may be a series of words or links that describe the cluster, but this cannot be searched in a way that is recognizable to the common user.

The metadata tag “solution” has its share of inescapable problems. The best metaphor for the meta tag solution is having a “catalog that is unique to each library”. First, there is not a standard format for metadata tags[5]. This means that understanding this information is almost impossible unless the author wrote these tags specifically for a particular web search engine. In which case it is not accessible to a broad number of client sites. Another problem with this method is that an author of a website is responsible for documenting the way in which the website is searched. This would be okay except lets face it computer scientists do not think like the average person. It is unlikely that the authors idea of how to search his website is a thorough enough description of the contents of his site.

Solution:

The prior attempts at a solution lead to the progression of a need for a more customizable solution to the problem. An automatic web categorization application that utilizes user defined categories and keywords would be a perfect solution to this problem. This application will allow a visitor to the website to navigate with ease through a vast number of websites. It will also narrow down the possible hits of a particular query or search using the user's profile of defined keywords and categories.

The first step in order to achieve this process is to evaluate an html document. Once a visitor has submitted a web address to be placed into the database. The html document from that website is retrieved. Once it is retrieved insignificant words such as "is", "the", and "and" are filtered out. Then each word retrieved has two identifiers associated with it. There is a count that is saved with the number of times the word occurred in the document and there is also a list of positions associated with each word. For instance given this sample document from www.christinavsbritney.com …

Christina vs. Britney

Britney is a hot chick. Britney Spears sings music. To visit the Britney website go to www.peeps.com/britney Christina Aguilera is a much better singer. To visit the Christina website go to www.christinaaguilera.com .

Stripping the insignificant words from the Document would give

Christina Britney Britney hot chick Britney Spears sings music Britney Christina Aguilera better singer Christina

The reason that the web addresses were dropped is because they were links and will be evaluated later. The significant word Description that would result from this document is as follows…

Word	Count	Position
Christina	**3**	**1, 11, 15**
Britney	**4**	**2,3,6,10**
Hot	**1**	**4**
Chick	**1**	**5**
Spears	**1**	**7**
Sings	**1**	**8**
Music	**1**	**9**
Aguilera	**1**	**12**
Better	**1**	**13**
Singer	**1**	**14**

Once the html is broken down into significant components it is then compared to the Category texts defined by the user of the categorization application. The categorization texts consist of words or phrases that are associated with a particular category.

For example if the user defined category states…

Category:

Music

Words:

Pop
Sings
Britney Spears
Christina Aguilera
Eminem

When this category is described by the user above is compared to www.britneyvschristina.com there are several exact matches. The exact match comparison is used to determine whether the website is in a particular category. The number n is the total number of exact matches and t is the total number of words evaluated in the document. The category score is derived from the equation…

Category Score = n/t

Christina	**3**	**1, 11, 15**
Britney	**4**	**2,3,6,10**
Hot	1	4
Chick	1	5
Spears	**1**	**7**
Sings	**1**	**8**
Music	**1**	**9**
Aguilera	**1**	**12**
Better	1	13
Singer	1	14

N = 11, T = 15

Category Score = 11/15

Thus the web page would most likely be placed in the category Music Category. Next in order to define whether or not a keyword will be directly associated with website let the definition be where the keyword k divided by total words t equal the keyword score. If the key words are together such as Britney Spears is derived by the absolute value of the closest word position of k1 – k2 divided by 1 divided by t times k1 plus k2 …

The Sum of (k1 – k2)(Count of k1 + Count of k2) all divided by T (Dr B I need help explaining)

Using the previous document the results for *Britney Spears* and *sings* are as follows

Sings Match = 1/15 =

Britney Spears Match = 4(|7-6|) + 1(|6-7|) = 5/1 = 5/15 = .333333

Thus Britney Spears would most likely be counted as an exact match where as Sings would not.

When this is finished the links are retrieved and www.peeps.com/britney and www.christinaaguilera.com undergo the same evaluation as stated above, thus finding more websites with similar interests. Therefore every time a single link is added in many more are found and evaluated.

Searching this database of information would be notably easier to navigate through than a database of clustered material with no background knowledge. This

would also allow the user to narrow his or her search engine to one of a specific area. An html document may be listed under many different categories. This allows the user of the application to modify and change the way the html is evaluated.

In order to update a changed website. All the user would have to do is run a re evaluation of the website in question. Or they may also run an overall evaluation of the categorization of the entire database. The process of categorization would be redone on the same sites automatically, thus a human would never have to go and visit the site at all instead this application would do all of the work for categorization based upon the preferences of the user. Thus creating a “cataloged web” library.

In order to create a visitor profile one must track the users links that he/she or she navigates throughout the web. Keeping track of the links clicked on would allow a creation of a master list of keywords the user wishes to view. For each link the user clicks a the keywords from that page are loaded into a user profile. By doing this when a user types in a query instead of searching the category information text, we search the list created by the tracking of the user and evaluate the pages returned again using his/ her profile. This will allow a second mechanism for filtering out sites that may not be what the user wanted, and adding sites of which the user may have wanted. This would allow for a more personalized search engine for each visitor of the site. Their “librarian” would be their past history of links they have chosen to click on as a source.

Works Cited

1. Callery, Anne, “Yahoo Cataloging the Web” 1996
2. Various, “Document categorization and Query Generation on the WWW Using Web Ace”, University of Minnesota
3. Modha and Spangler, “Clustering Hypertext with Applications to Web Searching”, IBM Almaden Research Center
4. Richmond Alan, “META Tagging for Search Engines”
5. Chris Welty Conversation October 2000, Vassar College Computer Science Department Poughkeepsie New York
6. Notess, Greg, “Internet search Techniques and Strategies” July 1997
7. Bizzozerom and Rana “Dynamic WAIS Book: an electronic book to publish and consult information distributed across a wide-area network” Environment Informatics Unit, Joint Research Center of the EC Ispra(VA), Italy
8. Koenemann and Belkin “A Case For Interaction: A Study of Interactive Information Retrieval Behavior and Effectiveness” Rutgers University New Bunswick New Jersey